Supplementary Information for
The Impact of U.S. County-Level Factors on COVID-19 Morbidity and Mortality.

Nevo Itzhak, Tomer Shahar, Robert Moskovich and Yuval Shahar

Yuval Shahar
Email: yshahar@bgu.ac.il

This PDF file includes:

Supplementary text
Figures S1 to S10
Tables S1 to S2
SI References

Data Sources

The data were extracted from several sources.

Daily reports of cumulative COVID-19 cases and deaths for each county and mask-wearing survey data (250,000 responses between July 2 and July 14, 2020) were taken from The New York Times data site [https://github.com/nytimes/covid-19-data]. Note the COVID-19 data used in this study have been collected up to November 28, 2020. Figure S1 presents the rate of COVID-19 cases [morbidity] (left) and COVID-19 induced deaths [mortality] (right), normalized by the county's population size, in each U.S. county. Figure S1 was created using the Plotly Express package for Python version 4.4.1 [https://plotly.com/python/plotly-express].

Epidemiological factors (diabetes, obesity and inactivity) data (2013) were obtained from the sites of the Centers for Disease Control and Prevention [https://www.cdc.gov].

Ethnicity, socioeconomic, educational attainment, population density and age distribution data (2018) were gathered from the U.S. Census Bureau site [https://www.census.gov].

The socioeconomic data included also the mode of transportation, sector of employment, and ethnicity, amongst other variables. The age dataset contained statistics for each age group featured.

Evaluation of the capacity of ICU, surgical ICU, coronary care unit and burn ICU beds data were taken from the Kaiser Health News site (KHN) [2]. The data were updated up to March 30, 2020 and do not include Veterans Affairs hospitals, since these hospitals do not file cost reports.

County-level presidential election results (in 2016) were taken from the Townhall site [https://townhall.com].

Other than the mask-wearing survey and the capacity of ICU beds, all U.S. county-level features were extracted from repositories created before the beginning of the spread of the COVID-19 pandemic.

The full features list, including their sources and descriptions, can be found in table S1.

Data Pre-Processing Methods

We merged the described datasets, using *Federal Information Processing Standard Publication* (FIPS) counties codes, and removed features that were highly correlated, either positively or negatively (absolute Pearson correlation coefficient greater than 0.9), using only one feature from each set of correlated features. There were no missing values in the final merged data set.

The features were normalized into a range of zero and one, aside from natural increasing rate, death rate, and total migration. These three features have negative values. Thus, we normalized these features into a range of minus one and one. Table S2 presents the features and their statistics: mean, standard deviation (STD), minimum (Min), the first quartile (Q1), median, the third quartile (Q3), and the maximum (Max). For example, to normalize the *Mean Commute Time* feature to have a value between zero and one, we divided it by 24 hours. This ensures the feature value will be in that range. Similarly, to normalize the *Mean Age* feature into to have a value between zero and one, we divided it by 120 years, since 120 is much greater than even the maximal value to be found for that feature.

Please note the statistics in table S2 are for the 3,071 counties that were used in the main paper and not all of the counties in the U.S.

Computational Methods

Figure S2 displays a histogram of the COVID-19 mortality rate and COVID-19 morbidity rate in the U.S. The top quartile thresholds were used to determine the class for each county for either COVID-19 morbidity or mortality.

There was no direct correlation between the morbidity and mortality rates in each county.
The computation of the SHAP values can be split into two stages:

1. Finding the optimal hyper-parameters (out of the combinations we tested) for our model using a Grid search with cross-validation. The chosen hyperparameters are the ones that resulted in the best average AUC, over the ten folds tested for each setting.
2. Using the determined optimal hyper-parameters to apply a stratified cross-validation procedure for computing the SHAP values for different subsets of the counties, eventually extracting SHAP values for all counties. The SHAP values found are for the specific fold that was used as a test set. For example, if 90% of the counties are used for training and the remaining 10% are used for testing, then we can only find SHAP values for those 10%. Thus, we must split the data into multiple folds, so that each county is a part of a test set once (allowing us to find the SHAP impact values for it). Therefore, each fold resulted in a unique model, and the overall SHAP values reflect an integration of all of the models.

It was also crucial to ascertain that indeed, each of those fold-specific models achieved a reasonable accuracy, otherwise one cannot assign importance to the determined SHAP values found. In our experiments, all models had a fairly high AUC and AUPRC scores making them meaningful.
It is also important to note that the procedure we used here is useful solely for finding a set of good classification models, in order to compute from them their respective SHAP values and determine the relative impact and direction of each of the features these models used, and is not

suitable for generating a good prediction model, which would typically entail using an additional hold-out data set on which the predicted optimal hyper-parameters would then be tested.

The combinations for the grid-search approach were a Cartesian product of the following parameters: Classifier with or without bootstrapping. The minimum number of samples required to split an internal node were 2, 3, 4 and maximal depth of 10, 15, 20, 25, 30. The minimum number of samples required to be at a leaf node were 3, 4, 5, 6 and the number of trees in the forest were 100 to 1000 with a step-size of 100.

For both models, we used a Random Forest algorithm to induce the classifier without bootstrapping.

The minimum number of samples required to split an internal node was 2, with a maximal depth of 15; and the random-generator seed was 2020.

For the mortality model, the minimum number of samples required to be at a leaf node was 5 and the number of trees in the forest was 200. For the morbidity model, the minimum number of samples required to be at a leaf node was 4 and the number of trees in the forest was 600. For the parameters we did not specify, we used the default. Random Forest classifier and Grid search approach was implemented with Python 3.7 Scikit-Learn [https://scikit-learn.org/] version 0.22.2. This algorithm is a way to reverse-engineer the output of any predictive algorithm. SHAP values are used when there is a complex model and we want to understand what decisions the model is making. In general, the SHAP values show how much a given feature changed our prediction and in which direction. SHAP was implemented using the SHAP package for Python version 0.35.0 [https://github.com/slundberg/shap].

Evaluation Metrics

This section describes the evaluation metrics we used, using as an example the distinction method we chose for the mortality rate; the morbidity rate was divided similarly.

The *True Positive Rate* (TPR) (also referred to as *sensitivity* or *recall*) is the proportion of counties who were classified by the model as being ”High Mortality” (true positives) out of counties who were labeled as being in the ”High Mortality” subset. The TPR represents the model’s ability to find all ”High Mortality” counties.

The *False Positive Rate* (FPR) (also referred to as the *specificity*) is the proportion of counties wrongly categorized as ”High Mortality” (false positives) out of the total number of labeled as ”Low Mortality”. The FPR represents the probability of a false alarm - i.e., incorrectly predicting a ”High Mortality” label.

The *Precision* (also referred to as the *Positive Predictive Value* (PPV) is the proportion of counties who were (correctly) labeled as belonging to the ”High Mortality” subgroup, out of those who were classified as belonging to the ”High Mortality” subgroup. The precision is the model’s accuracy when predicting the ”High Mortality” label for a given county.

The *Receiver Operating Characteristic* (ROC) graph displays on the vertical axis the TPR, and on the horizontal axis the FPR, for all of the binary classification thresholds.

The *Precision-Recall* graph displays on the vertical axis the recall (i.e., the TPR) and on the horizontal axis the precision (PPV) for different probability thresholds.

To quantify the classification performance of both the morbidity and mortality models, we used two evaluation metrics: (1) the *Area Under the ROC curve* (AUROC), and (2) the A*rea Under the Precision-Recall Curve* (AUPRC). Higher AUROC and AUPRC values imply that the model is more accurate. Interested readers might wish to read more about these metrics and the relationship between them elsewhere [1].

Results

Figures S3 and S4 present the twenty most impactful features (vertical axis) on the COVID-19 morbidity and mortality model, versus the rate of COVID-19 cases and deaths (horizontal axis), respectively.

Figure S3 present a variable (vertical axises) versus the rate COVID-19 cases (horizontal axises) in U.S. Orange points represent counties with high COVID19 morbidity, and blue points represent counties with low COVID-19 morbidity counties. The variables order follows the order presented in the results figure in the paper. For example, the upper left variable (i.e., *Mean Commute Time*) is the most impactful factor in the morbidity model followed by the feature *Always Wears a Mask*. The feature *Total Migration* is the sixth most impactful feature and *Median Age* is the fourteenth most impactful feature. As can be seen from the upper left plot, *Mean Commute Time* versus COVID-19 cases, there is a slight negative connection between the *Mean Commute Time* and COVID-19 cases. This negative correlation fits with the results presented in the paper, namely that high values of *Mean Commute Time* led to lower COVID-19 morbidity rates in a county.

The same semantics are present for figure S4 with COVID-19 mortality. For example, the upper left variable (i.e., *African-American Population*) is the most impactful factor in the mortality model, followed by the feature *Caucasian Population*. The feature *Physical Inactivity Level* is the sixth most impactful feature, and *Hispanic Population* is the fourteen-the most impactful feature. As can be seen from the upper left plot, *African-American Population* feature's plot versus the COVID-19 deaths, there is a slightly positive connection between the *African-American Population* feature and the COVID-19 deaths. This positive correlation fits in with the results presented in the paper, namely that high values of the *African-American Population* feature led to higher COVID-19 morality rates in a county.

Behavior of Morbidity and Mortality Predictors Over Time. Some features seem to change over time, both their absolute impact and their *direction* of influence. We analyzed all features correlated with morbidity and mortality from April 1, 2020, until November 28, 2020. Figure S5 contains eighteen more Pearson correlation plots, over time, between variables for each model. This figure is provided in addition to the ten plots provided in the paper.

For example, in figure S5.c, in early August 2020, the value of *Total Poverty Rate* feature had a high positive (+0.35) impact on the number of COVID-19 cases. However, when calculated using data collected up to November 28, 2020, the correlation is almost twice less (+0.15).

Urbanity and Ruralness Of Effected Counties Over Time. The total ”*Rural-Urban Continuum Codes*” (RUCC) distribution can be seen in figure S6. The USDA assigns each county in the U.S. an RUCC value from 1 to 9, representing how urban or rural it is based on the county’s own population and the population of adjacent counties.

An RUCC of one represents the most ”urban” counties (metro areas with a population of 1 million or more); an RUCC of nine represents the most ”rural” counties (less than 2,500 population, not adjacent to a metro area). Typically, counties with an RUCC of 1 to 3 are considered metropolitan and comprise approximately 37% of the counties in the U.S., while counties with an RUCC of 4 to 9 are considered non-metropolitan and comprise 63% of the counties in the U.S.

Figure S7 presents the distribution of RUCC values between counties in the top quartile for the 32 different variables value. The horizontal axis denotes the number of counties, the vertical axis the RUCC value. For example, the upper left plot in figure S7 presents the histogram of the top quartile counties of *Mean Commute Time* values. As can be seen from this histogram, the top quartile counties of *Mean Commute Time* values are more metropolitan counties, which is reasonable. Similar behavior can be found in other features, such as *Always Wears a Mask, Population Density, Adults with Academic Education, Median Household Income*.

Effect of the Same factors at an Individual Level versus at a Population Level. Several features which are known to increase the mortality from COVID-19 (such as older age, gender, and ICU bed availability) were considered in our study but did *not* appear in the list of highly impacting factors, or appeared with a low impact weight. In particular, the low impact of the *age* features on morbidity and mortality in our results might initially seem surprising, considering the well known association in COVID-19 individual patients between being at an elderly age and suffering the most severe complications; and the higher propensity for death in males.

The likely reason for this lack of association in our current study is a *low variance of these features among counties*, as opposed to their high variance among individuals.

Figure S8 presents box plots for the various age brackets used in our models. As can be seen, there is little variance between almost all of these features. There are a few outliers, such as population in the 20-29 and 70+ bracket. Notice that the quartiles are spread out rather evenly.

In contrast, figure S9 presents box plots for the percentage of Caucasian and African-American populations across counties. We can see a very high variance inside each feature, and a large difference between these features themselves. This is likely to be the reason that age did not emerge in our methodology as one of the top predictive features for morbidity or mortality: for an *individual*, age is a highly significant predictor of the outcome. But when predicting the effect of COVID-19 for a whole *county*, the age distribution withinn the county is not a very useful predictor, because the distribution of age values is quite similar between different counties.

As discussed also in the main paper, the Pearson correlation between obesity and diabetes prevalence is relatively high (0.698), they impacted the model differently. Figure S10 presents the box plots of *Diabetes* and *Obesity* features. Diabetes prevalence has low variance ($5.76*10^{-4}$) between counties, leading to relatively random results when we compute the SHAP values. In contrast, the variance in obesity prevalence was much higher ($2.02 * 10^{-3}$).

Figures

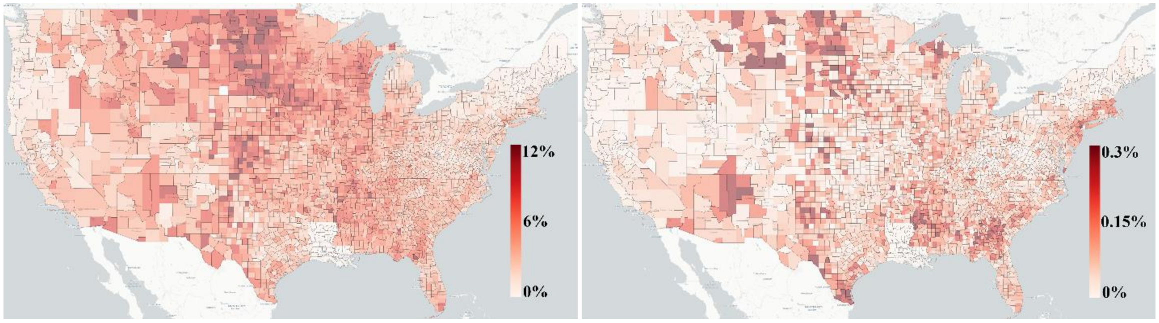


Fig. S1. Rate of COVID-19 cases [morbidity] (left) and COVID-19 induced deaths [mortality] (right), normalized by the county's population size. The color represents the rate of morbidity or mortality, respectively. COVID-19 data have been collected up to November 28, 2020.

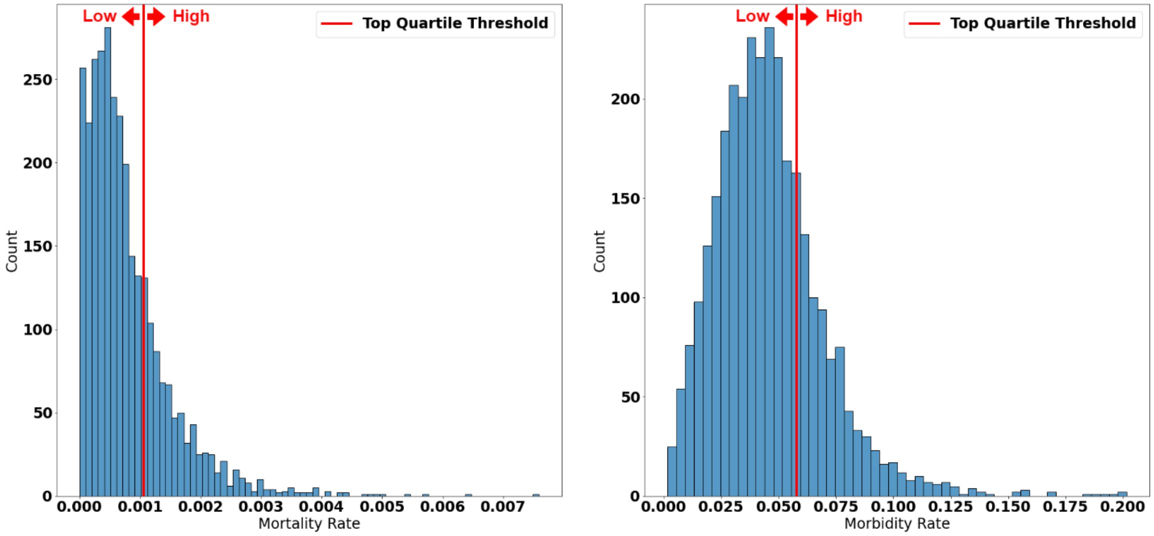
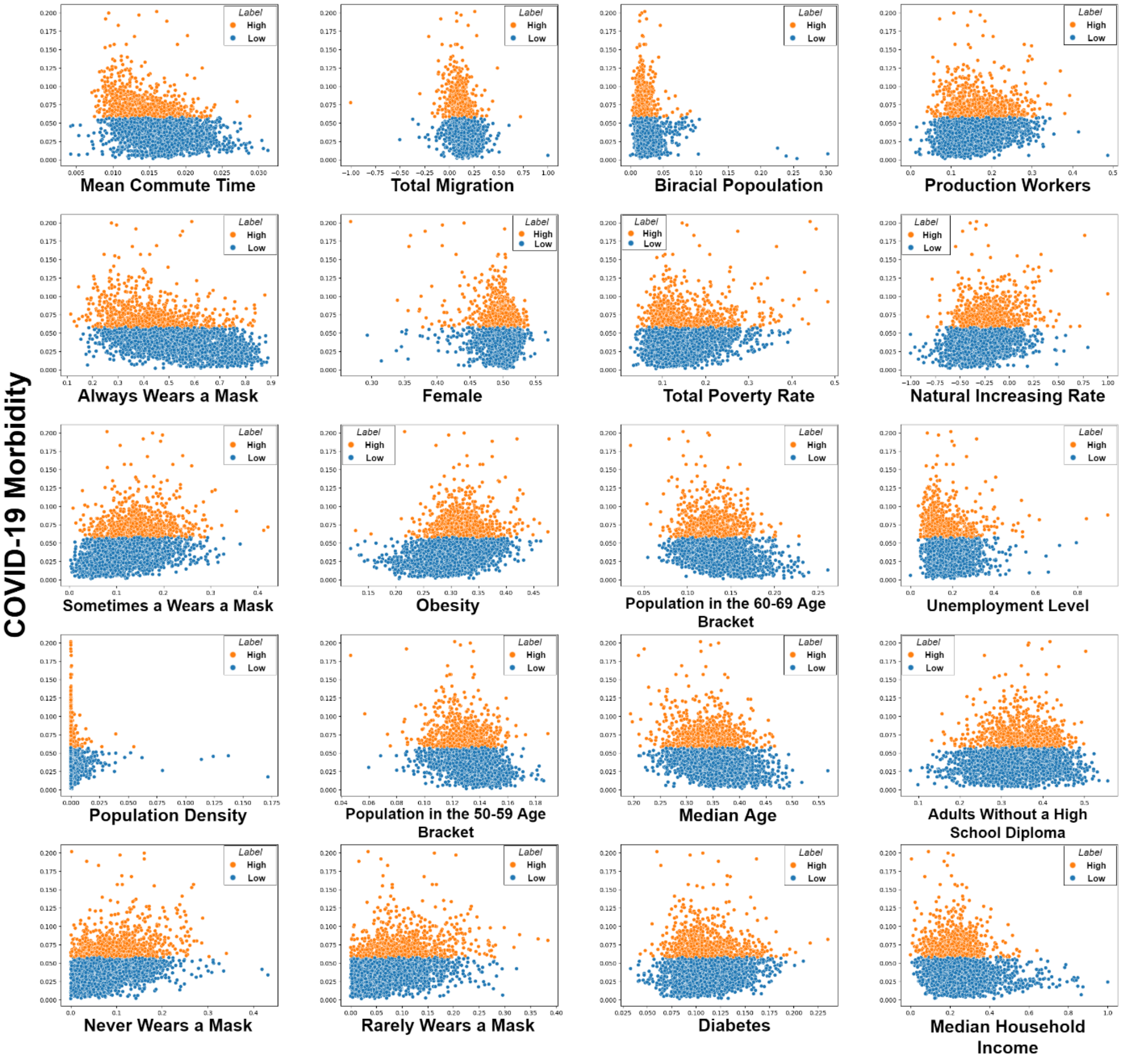


Fig. S2. Histogram of the COVID-19 mortality rate (left) and COVID-19 morbidity rate (right). The red vertical line represents the top quartile thresholds.



Mean Commute Time

Total Migration

Biracial Popoulation

Production Workers

Always Wears a Mask

Female

Total Poverty Rate

Natural Increasing Rate

Sometimes a Wears a Mask

Obesity

Population in the 60-69 Age Bracket

Unemployment Level

Population Density

Population in the 50-59 Age Bracket

Median Age

Adults Without a High School Diploma

Never Wears a Mask

Rarely Wears a Mask

Diabetes

Median Household Income

Fig. S3. The twenty most impactful features on the COVID-19 morbidity model. Each graph represents the variable values (vertical axis) versus the COVID morbidity rate (horizontal axis) in each county. Orange points represent counties with high COVID-19 morbidity and blue points represent counties with low COVID-19 morbidity counties.

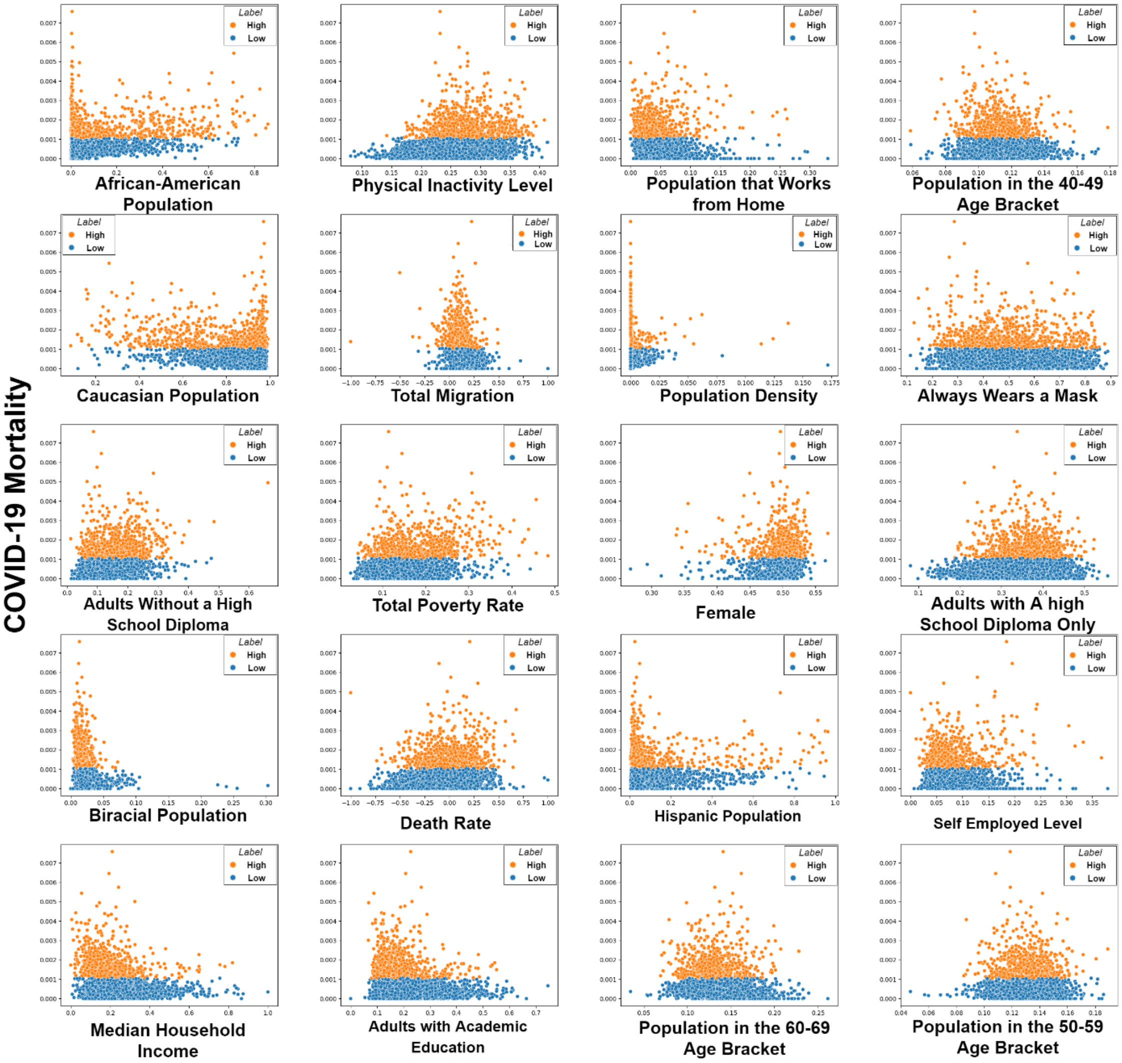


Fig. S4. The twenty most impactful features on the COVID-19 mortality model. Each graph represents the variable values (vertical axis) versus the COVID mortality rate (horizontal axis) in each county. Orange points represent counties with high COVID-19 mortality and blue points represent counties with low COVID-19 mortality counties.

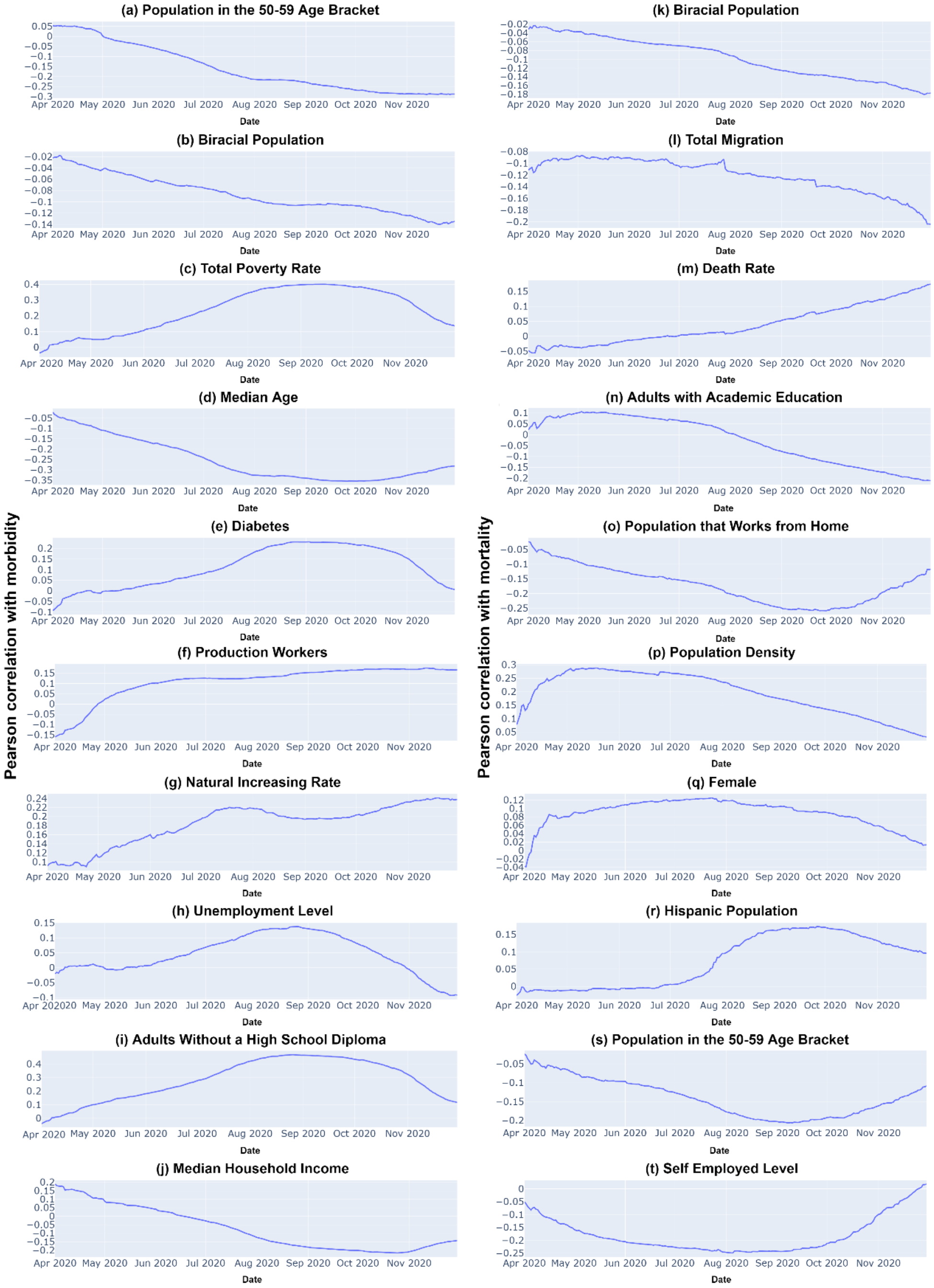


Fig. S5. The Pearson correlations between an additional nine features (outside of the ones shown in the main paper) for each model and the percentage of morbidity (a through j) and mortality (k through t). The features are sorted such that the more impactful ones are higher (i.e., a, j have more impact on the morbidity and mortality model than b, l, respectively). The correlation is plotted over time, from the 1st of April until the 28th of November.

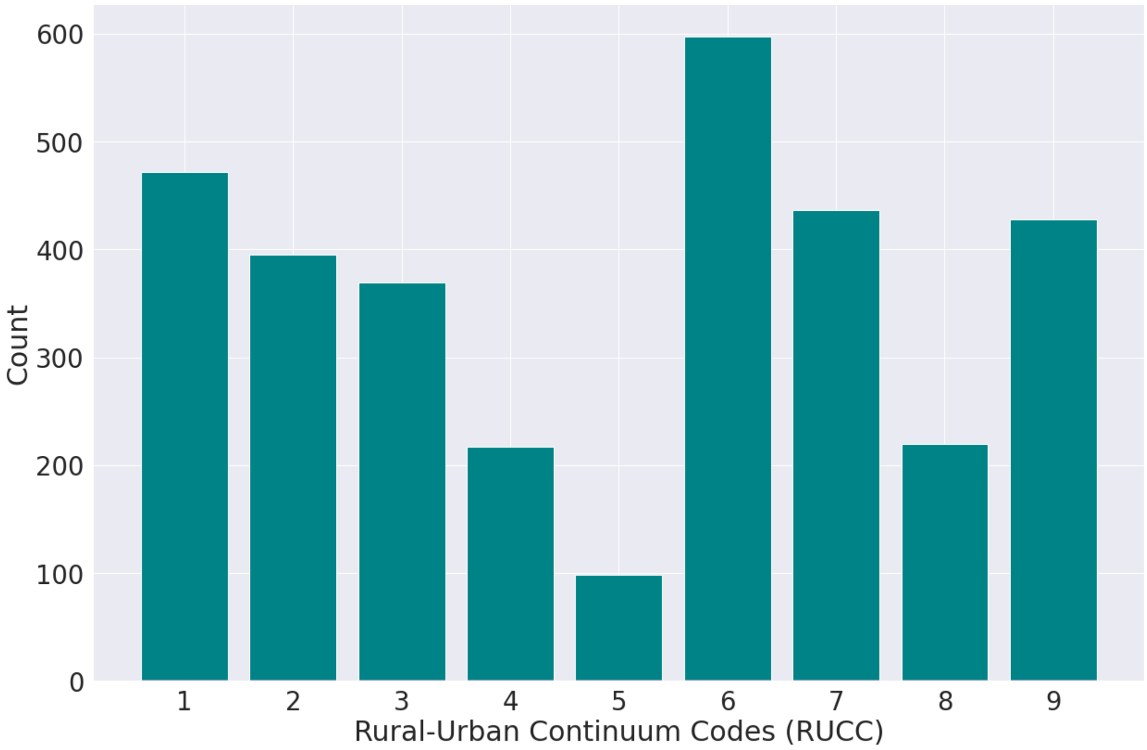


Fig. S6. Distribution of Rural-Urban Continuum Codes (RUCC) values in the U.S. The horizontal axis denotes the number of counties, the vertical axis the RUCC value.

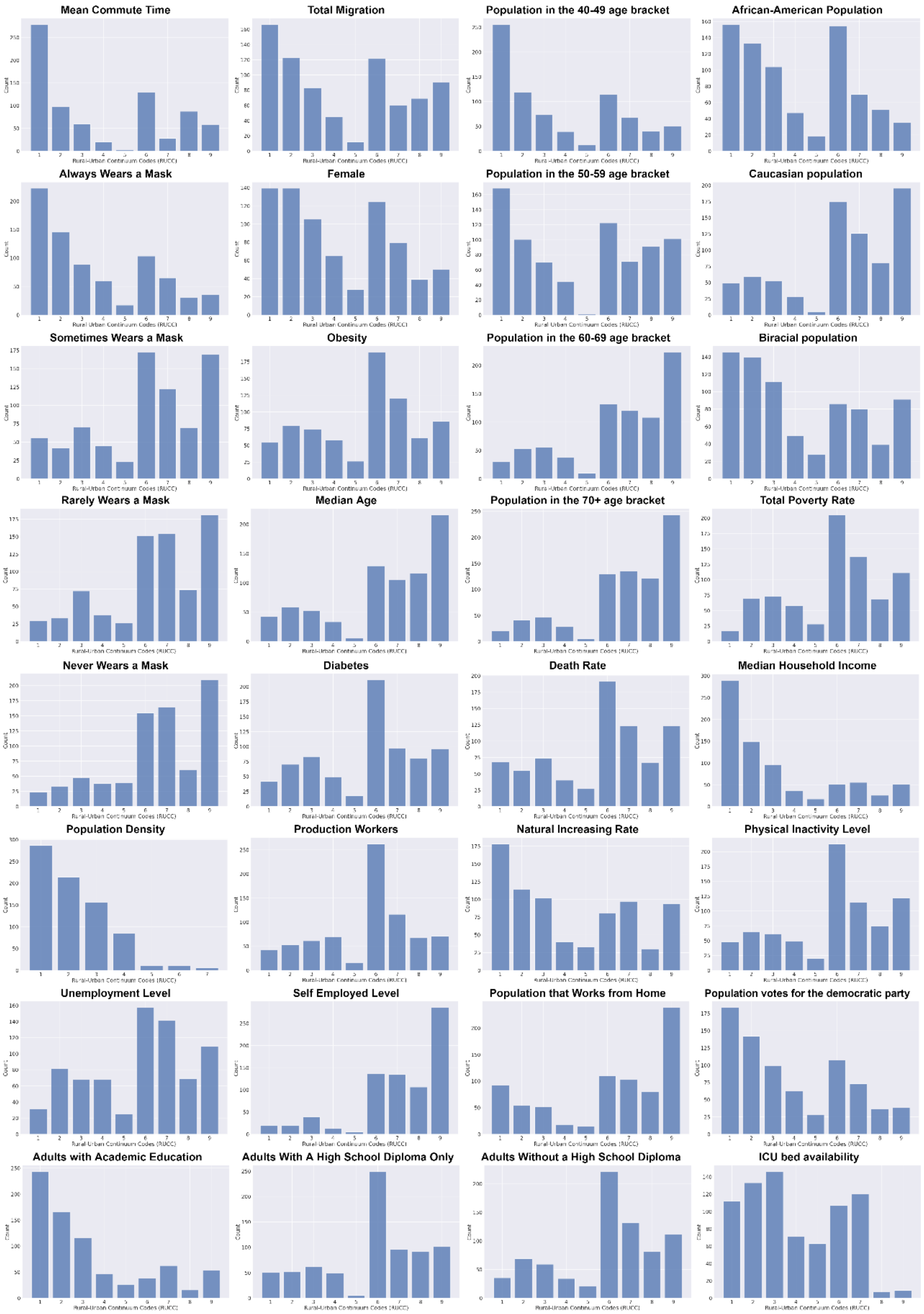


Fig. S7. Distribution of Rural-Urban Continuum Codes (RUCC) values between counties that are in the top quartile for the different variables value. The horizontal axis denotes the number of counties, the vertical axis the RUCC value.

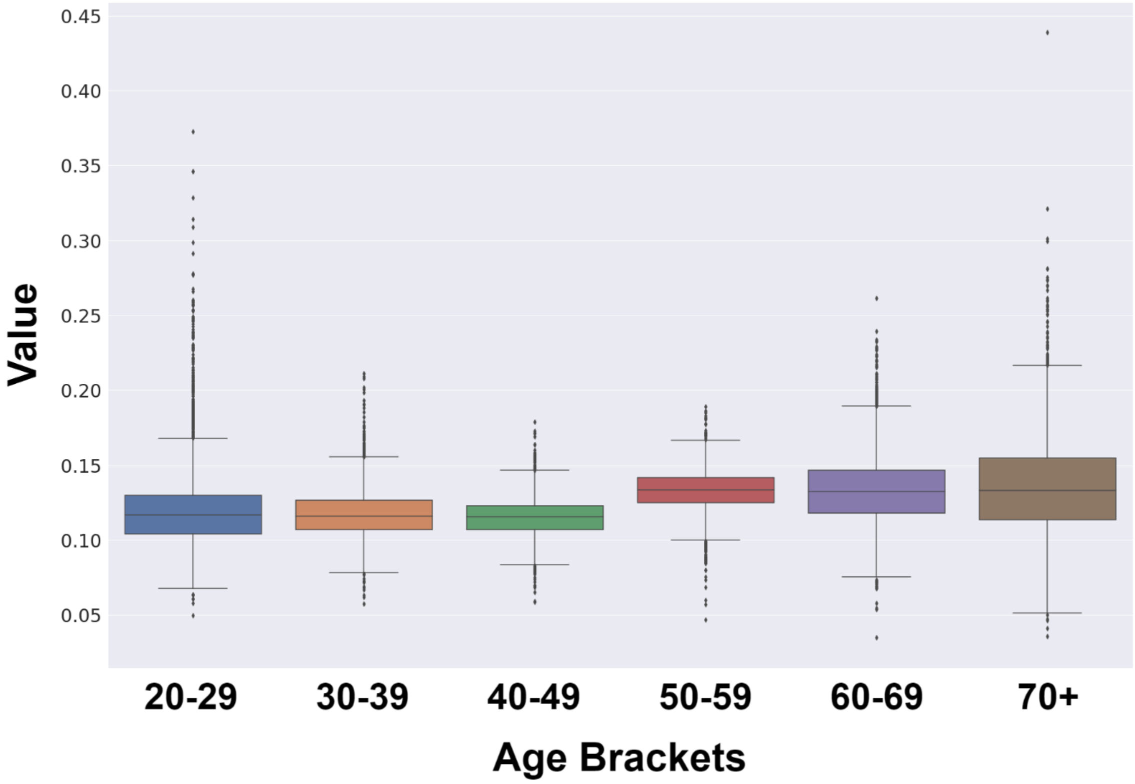


Fig. S8. Box plots for African-American population percentage and for Caucasians' percentage across counties.

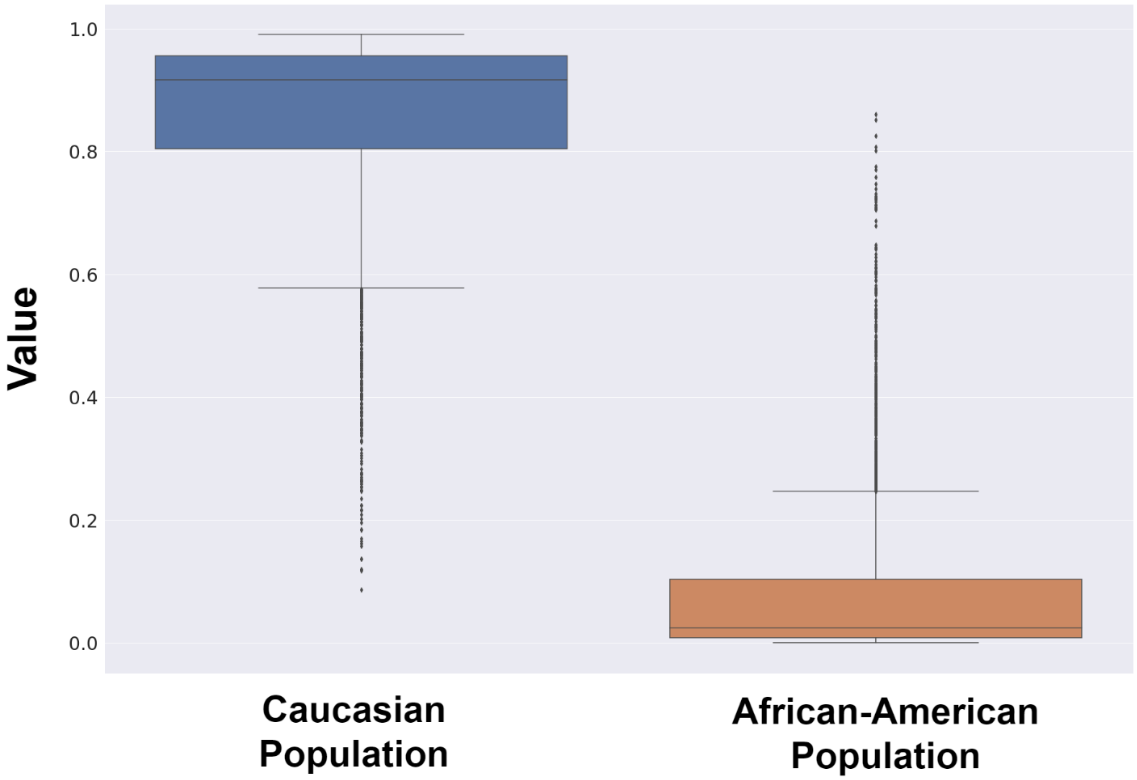


Fig. S9. Box plots for African-American population percentage and for Caucasians' percentage across counties.

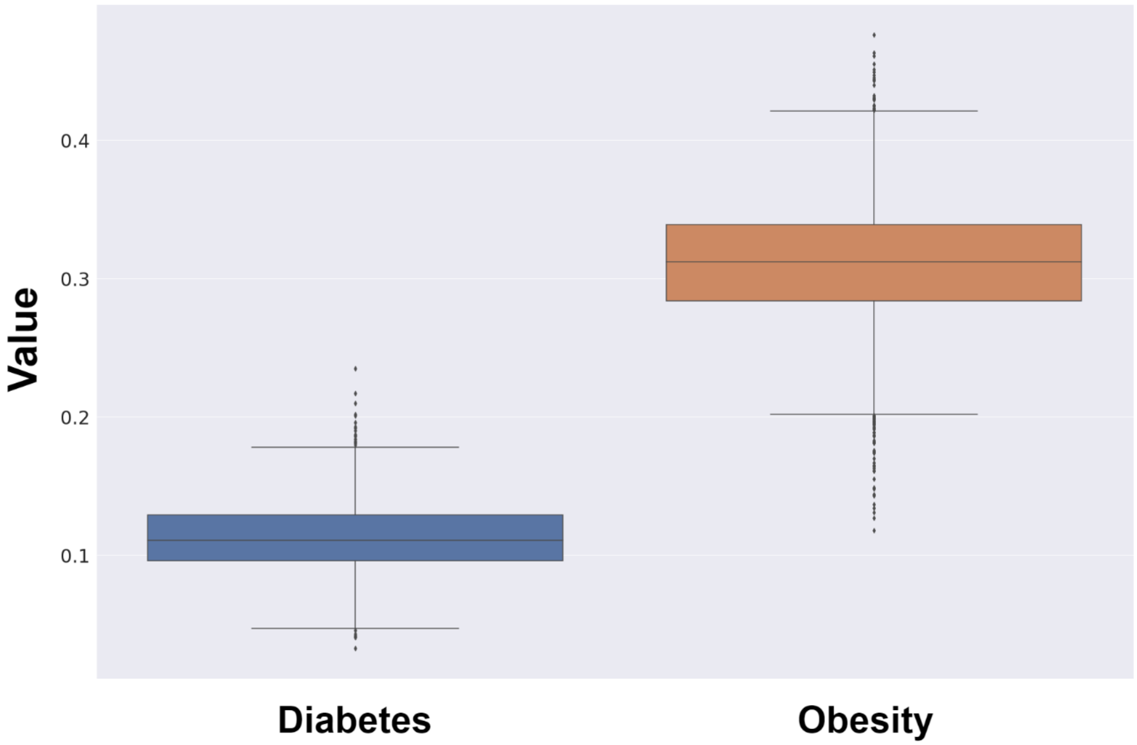


Fig. S10. Box plots for Diabetes and Obesity percentage across counties.

Tables

Table S1. All features used, their source and description.

Feature Name	Source	Description
Percent of Adults with Less than a High School Diploma	United States Department of Agriculture (USDA), Economic Research Service	Percent of Adults with Less than a High School Diploma
Percent of Adults with a High School Diploma only		Percent of Adults with a High School Diploma only
Percent of adults that completed some College or Associates Degree		Percent of adults that completed some College or Associates Degree
Percent of Adults with a Bachelors Degree or Higher		Percent of Adults with a Bachelors Degree or Higher
Total Poverty Rate		Percent of population that are impoverished
Death Rate		Percent of population that died in 2019
Natural Increasing Rate		Increase in percentage of population size in 2019
Total Migration		Number of people immigrating/emmigrating to the county as a percentage of total population
Unemployment Level		Percentage of unemployed people
Median Household Income		Median income of county
Diabetes	Living Atlas of the World - Diabetes, Obesity, and Inactivity by US County	Percentage of people with diabetes in the county
Physical Inactivity Level		Percentage of people that do not perform physical activity
Obesity		Percentage of obese people in the county
Female	United States Census Bureau	Percentage of females in the county
African-American Population		Percentage of African-Americans in the county
Caucasian Population		Percentage of Caucasians in the county
Asian Population		Percentage of Asians in the county

Hispanic Population		Percentage of Hispanics in the county
American Indian Population		Percentage of Indians in the county
Biracial Population		Percentage of people with mixed heritage in the county
ICU Beds per person	Kaiser Health News (KHN)	Number of beds in ICU wards per person
Percentage of Democratic Voters	https://townhall.com/	Percent of people that voted for the Democrats in the 2016 elections
Voting Difference		Absolute difference between percent of Democratic and Republican voters for the 2016 elections
Voting Turnout		Percent of eligible voters in the 2016 elections (out of the whole population)
Professional Workers	United States Census Bureau	Percent of population that provide professional services
Service Workers		Percent of population that provide government services
Office Workers		Percent of population that work in offices in the private sector
Construction Workers		Percent of population that work in construction work
Production Workers		Percent of population that work in production facilities (e.g., factories)
Drive to Work		Percent of population that commutes by driving
Carpools to Work		Percent of population that commutes by carpooling
Public Transportation to Work		Percent of population that commutes with public transportation

Walk to Work		Percent of population that commutes by walking
Other Transportation to Work		Percent of population that commutes by other means
Population that Works from Home		Percent of population that works from home
Mean Commute Time		Average amount of time it takes people to get to work (in hours)
Private Work Sector		Percent of people that work in the private sector
Public Work Sector		Percent of people that work in the public sector
Self Employed		Percent of people that are self-employed
Family Work Sector		Percent of people that work in a family business
Never Wears a Mask	New York Times Github Repository	Percent of people that never wear a mask
Rarely Wears a Mask		Percent of people that rarely wear a mask
Sometimes Wears a Mask		Percent of people that sometimes wear a mask
Frequently Wears a Mask		Percent of people that frequently wear a mask
Always Wears a Mask		Percent of people that always wear a mask
Population Density	United States Census Bureau	Number of people per square mile in the county
Median Age	United States Department of Agriculture (USDA), Economic Research Service	Median age of people in the county
Population in the 20-29 age bracket		Percent of population between the ages of 20-29
Population in the 30-39 age bracket		Percent of population between the ages of 30-39
Population in the 40-49 age bracket		Percent of population between the ages of 40-49
Population in the 50-59 age bracket		Percent of population between the ages of 50-59
Population in the 60-69 age bracket		Percent of population between the ages of 60-69
Population in the 70+ age bracket		Percent of population over the age of 70

Table S2. Descriptive statistics for all features: mean, standard deviation (STD), minimum (Min), the first quartile (Q1), median, the third quartile (Q3), and the maximum (Max).

Feature Name	Mean	STD	Min	Q1	Median	Q3	Max
Percent of Adults with Less than a High School Diploma	0.134	0.063	0.012	0.088	0.121	0.172	0.663
Percent of Adults with a High School Diploma only	0.344	0.071	0.081	0.299	0.347	0.394	0.556
Percent of adults that completed some College or Associates Degree	0.308	0.052	0.058	0.273	0.307	0.342	0.573
Percent of Adults with a bachelor's degree or Higher	0.214	0.092	0	0.15	0.191	0.254	0.746
Total Poverty Rate	0.151	0.061	0.026	0.108	0.141	0.182	0.484
Death Rate	-0.055	0.238	-1	-0.201	-0.048	0.107	1
Natural Increasing Rate	-0.261	0.204	-1	-0.393	-0.273	-0.147	1
Total Migration	0.137	0.085	-1	0.098	0.134	0.176	1
Unemployment Level	0.176	0.076	0	0.124	0.161	0.21	0.946
Median Household Income	0.237	0.119	0	0.159	0.219	0.289	1
Diabetes	0.113	0.025	0.033	0.096	0.111	0.129	0.235
Physical Inactivity Level	0.26	0.052	0.081	0.227	0.259	0.295	0.414
Obesity	0.311	0.045	0.118	0.284	0.312	0.339	0.476
Female	0.499	0.022	0.268	0.494	0.503	0.51	0.569
African-American Population	0.091	0.143	0	0.009	0.025	0.104	0.861
Caucasian Population	0.851	0.157	0.087	0.804	0.916	0.956	0.99
Asian Population	0.015	0.027	0	0.005	0.007	0.014	0.43
Hispanic Population	0.096	0.139	0.006	0.024	0.043	0.1	0.964
American Indian Population	0.021	0.067	0	0.004	0.006	0.013	0.857
Biracial Population	0.021	0.014	0	0.013	0.018	0.024	0.303
ICU Beds per person	10^{-4}	$5*10^{-3}$	0	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$	0.028
Percentage of Democratic Voters	0.313	0.149	0.031	0.204	0.283	0.394	0.893
Voting Difference	0.392	0.207	0	0.224	0.405	0.554	0.916
Voting Turnout	0.75	0.052	0.457	0.732	0.76	0.782	0.907
Professional Workers	0.314	0.064	0.114	0.273	0.305	0.349	0.69
Service Workers	0.181	0.036	0	0.157	0.177	0.2	0.464
Office Workers	0.218	0.03	0.048	0.199	0.22	0.238	0.372
Construction Workers	0.127	0.041	0.033	0.099	0.122	0.149	0.364
Production Workers	0.16	0.058	0	0.118	0.156	0.197	0.487
Drive to Work	0.799	0.065	0.185	0.774	0.81	0.84	0.972
Carpools to Work	0.099	0.029	0	0.081	0.095	0.113	0.293
Public Transportation to Work	0.008	0.021	0	0.001	0.003	0.007	0.424
Walk to Work	0.03	0.031	0	0.014	0.022	0.038	0.497
Other Transportation to Work	0.015	0.012	0	0.008	0.013	0.019	0.211
Population that Works from Home	0.048	0.031	0	0.029	0.041	0.058	0.33
Mean Commute Time	0.016	0.004	0.004	0.014	0.016	0.019	0.031
Private Work Sector	0.752	0.073	0.321	0.718	0.763	0.803	0.888

Public Work Sector	0.167	0.059	0.044	0.126	0.156	0.193	0.623
Self Employed	0.078	0.039	0	0.053	0.068	0.092	0.38
Family Work Sector	0.003	0.005	0	0.001	0.002	0.003	0.08
Never Wears a Mask	0.081	0.059	0	0.034	0.069	0.115	0.432
Rarely Wears a Mask	0.084	0.056	0	0.041	0.074	0.116	0.384
Sometimes Wears a Mask	0.122	0.058	0.003	0.08	0.116	0.157	0.422
Frequently Wears a Mask	0.207	0.062	0.029	0.164	0.204	0.247	0.549
Always Wears a Mask	0.506	0.152	0.115	0.392	0.496	0.611	0.889
Population Density	0.002	0.007	0	0	0	0.001	0.172
Median Age	0.349	0.044	0.194	0.322	0.348	0.373	0.568
Population in the 20-29 age bracket	0.121	0.03	0.05	0.104	0.117	0.13	0.373
Population in the 30-39 age bracket	0.118	0.017	0.058	0.107	0.116	0.126	0.211
Population in the 40-49 age bracket	0.115	0.013	0.059	0.107	0.115	0.123	0.179
Population in the 50-59 age bracket	0.133	0.014	0.047	0.125	0.134	0.142	0.189
Population in the 60-69 age bracket	0.134	0.025	0.035	0.118	0.132	0.147	0.261
Population in the 70+ age bracket	0.136	0.035	0.036	0.113	0.133	0.155	0.439

SI References

1. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning. pp. 233–240 (2006)
2. Schulte, F., Lucas, E., Rau, J., Szabo, L., Hancock, J.: Millions of older americans live in counties with no icu beds as pandemic intensifies. Kaiser Health News (2020)